表現序列標幟(EST)功能註解分析的新流程

邱怡嘉 [1]、魏夢麗 [2]、呂椿棠 [2]、陳政道 [2]、呂秀英 [2]*
[1] 私立朝陽科技大學生物技術研究所
[2] 行政院農委會農業試驗所作物組

摘要

表現序列標幟（expressed sequence tag, EST）是 cDNA 序列片段，直接研究 EST 可獲取基因表現訊息。針對 EST 功能的註解、分類及表現量統計分析，本研究建立了一套完整且快速的新流程：將 EST 資料經 CAP3 軟體聚集和接合後所獲得的較高品質和去除多餘性的單一序列(unique sequences)，去比對 TIGR(The Institute for Gene Research)之 TC(tentative consensus sequence)資料庫以進行功能註解，再藉由 TC 搜尋結果連結到 GO(Gene Ontology) 與 MIPS(Munich Information Center for Protein Sequences)之功能分類系統進行功能分類，最後進行基因功能表現量之統計。褐飛蝨[*Nilaparvata lugens* (Stal.)] 是亞洲地區重要的水稻(*Oryza sativa* L.)蟲害，為利於瞭解水稻對褐飛蝨的基因防禦機制，本研究以取自 NCBI 之 dbEST 資料庫中受褐飛蝨誘導的 EST 序列資料作為應用材料，利用所建立的功能註解分析流程，進行不同水稻抗、感性品種間基因功能表現的差異性比較。分析結果顯示，受褐飛蝨誘導之水稻抗、感品種在 GO 功能分類上，皆以生物程序的功能表現為主；而在更細分的 MIPS 功能分類上，抗性品種在生物程序中的代謝相關基因的表現量較感性品種為高，同時抗性品種啟動了感性品種所沒有的細胞逆境防禦反應(援救、防禦和毒性)、轉錄、及細胞生長和形態發生的基因功能表現。

關鍵詞： 表現序列標幟、聚集及接合、功能註解、分類目錄、基因表現量統計、水稻、褐飛蝨。

* 通信作者, iying@wufeng.tari.gov.tw
投稿日期：2006 年 8 月 4 日
接受日期：2006 年 12 月 6 日
作物、環境與生物資訊 4:49-64 (2007)
Crop, Environment & Bioinformatics 4:49-64 (2007)
189 Chung-Cheng Rd., Wufeng, Taichung Hsien 41301, Taiwan ROC

A New Procedure of Functional Annotation and Analysis in Expressed Sequence Tag (EST)

Yi-Chia Chiu[1], Meng-Li Wei[2], Chun-Tang Lu[2], Cheng-Tao Chen[2] and Hsiu-Ying Lu[2]*

[1] Graduate Institute of Biotechnology, Chaoyang University of Technology, Wufeng, Taichung Hsien 41349, Taiwan ROC

[2] Crop Science Division, Agricultural Research Institute, Wufeng, Taichung Hsien 41301, Taiwan ROC

ABSTRACT

Expressed sequence tag (EST) is the fragment of cDNA sequence. Direct study of EST is therefore helpful to obtain the gene expression information. A complete and rapid procedure including EST functional annotation, classification and representation statistics was established as the followings. The first step was to use unique sequences of high quality from removing redundancy after clustering and assembling by CAP3 software to homology search TC (tentative consensus sequence) database of TIGR (The Institute of Genomic Research). The second step was to link the functional catalogs of GO (Gene Ontology) and MIPS (Munich Information Center

for Protein Sequences) from tentative annotation of the selected TC sequences. Finally the statistics of gene functional representation was calculated. Brown planthopper [*Nilaparvata lugens* (Stal.)] is the major pest of rice (*Oryza sativa* L.) in Asia. In order to know the mechanisms of rice genes resistant to brown planthopper, we used rice EST sequences induced by brown planthopper from NCBI-dbEST database as materials to compare the gene expression differences between the resistant and sensitive varieties. Using our procedure of EST functional annotation and analysis, the results showed that the dominant gene function of rice ESTs induced by brown planthopper was biological process based on GO functional category. According to minute classification of MIPS functional category system, the genes of resistant rice variety expressed richer metabolism function than that of sensitive rice variety. The resistant rice variety switched on the gene functions related to cell rescue/defense and virulence, transcription, and cell growth/morphogenesis, but sensitive variety did not.

Key words: Expressed sequence tag, Clustering and assembly, Functional annotation, Classification catalogue, Statistics of gene representation, Rice, Brown planthopper.

前言

基因組研究對於鑑定基因表現，廣泛應用表現序列標幟(expressed sequence tag, EST)來探討不同組織的基因表現(Adams *et al*. 1991, Rudd 2003, Wei *et al*. 2005a, 2005b)。EST 的產生是由一特定生長發育期或特定生理狀態下之組織將其表現的訊息RNA(messenger RNA, mRNA)分離出，mRNA 經反轉錄酶(reverse transcriptase)產生互補 DNA (complementary DNA, cDNA)，將其載入寄主細胞產生許多 cDNA 副本，這些寄主細胞包含 cDNA 的族群稱為 cDNA 基因庫(library)，再經由 cDNA 基因庫挑選殖體(clone)，以有系統的方式由 cDNA 的 5'端或 3'端進行單一自動化定序(single-pass automatic sequencing)所得到的序列，通常長度是 200～500 bp (Adams *et al*. 1991)。EST 雖然只是 cDNA 的片段序列，但具快速、低廉且直觀的優點，故成為連接結構基因組學和功能基因組學的一道橋樑，進行 EST 分析可提供基因辨識、新基因發現及基因表現差異性等研究，是最快速獲得序列資訊及提供功能性基因組研究之來源(Kelley *et al*. 1991, Adams *et al*. 1992, 1993, Rounsley *et al*. 1996, Rudd 2003, Wei *et al*. 2005a,b)。對於禾穀類作物而言，EST 資料可提供讀取基因的重要來源，以作為基因組圖譜定位及完整基因表現分析的基礎，更期望能透過 EST 來闡明功能性基因組及二級代謝物的生成及調節途徑，進而了解植物與生態環境間的關係(Sreenivasuls *et al*. 2002, Varshney *et al*. 2002, Wei *et al*. 2005a)。目前全世界研究單位都在針對禾穀類作物或其他重要作物進行大規模的 EST 計畫，各 EST 資料庫正快速增加資料中(Kantety *et al*. 2002)。獲得各種核酸和蛋白質序列的目的在於瞭解這個序列在生物體中扮演了什麼功能，此過程稱為功能註解(functional annotation)(Kawai *et al*. 2001)。透過已註解的 EST 序列，就可以進行其上下游資訊的全面分析，包括新基因發現、重複序列測定、調控基因確定、完整基因分析等(Jongeneel 2000, Wolfsberg and Landsman 2001, Sreenivasuls *et al*. 2002, Wei *et al*. 2005a)。

一般針對 EST 功能註解的方法，通常是直接利用 BLAST (basic local alignment search tool) 去比對 NCBI 資料庫 (http://www.ncbi.nih.gov/BLAST/) 或 MIPS 資料庫(MAtDB, http://mips.gsf.de/proj/plant/jsf/athal/searchjsp/searchSequence.jsp)，然後再將所比對到的相似性序列進一步利用 MIPS 功能分類系統(MIPS Functional Catalogue, MIPS_FunCat；http://mips.gsf.de/proj/funcatDB/search

_main_frame.html)來進行功能分類(Lai *et al*. 2004; Shim *et al*. 2004; Brenner *et al*. 2005)。MIPS_FunCat 是由慕尼黑蛋白質序列訊息中心(Munich Information Center for Protein Sequences, MIPS)所維護，主要是提供生物結構和組織之分類系統，以註解原核生物、單細胞真核生物、植物和動物之蛋白質在細胞內所扮演的角色(Frishman *et al*. 2001)，內含 22 個主要功能種類所組成的目錄，且在主要功能種類下又有再細分的功能分類，可用來對不同生物之基因組進行註解之分類，是研究者對基因及其產物進行功能分類時常參考的資料庫(Andreas *et al*. 2004, Mewes *et al*. 2006)。但是直接去對 NCBI 資料庫進行相似性序列比對，在所獲得之結果當中，往往比對到之相似性序列的來源物種包含很多，必須花相當多的時間去逐一檢視並過濾所比對到之相似性序列，來篩選出屬於查詢物種的序列。而 MIPS_MAtDB 的現有版本(2.0)資料庫內雖然也納入部分的水稻基因組序列元件，但仍以其主要研究計畫目標之阿拉伯芥(*Arabidopsis thaliana*)基因組序列為主。在搜尋到相似性序列後，若欲進而利用 MIPS_FunCat 來進行功能分類，則發現必須根據所獲得之相似性序列之功能註解的描述文字擷取關鍵字，直接於 MIPS_FunCat 的網頁畫面中輸入查詢文字，才能比對到其相關的功能類別。一旦關鍵字不同，所獲得之分類結果亦不相同，又關鍵字指定不夠，也會對應到多個功能類別，故難以確知真正的功能分類為何。由此可見，藉由比對 NCBI 或 MIPS_MAtDB，再直接利用 MIPS_FunCat 進行 EST 的功能註解與分類，除非經由一再核對及篩選的反覆過程，否則無法迅速且準確地獲知其真正的功能，此將嚴重影響整體分析的進度，故有必要利用其他途徑或去比對其他資料庫來建立一個適合於 EST 資料之功能註解的分析流程。

水稻(*Oryza sativa* L.)是世界上最主要的糧食作物之一，在許多國家均有栽培，尤其在亞洲有超過 90%的產量及消耗量，且有 1/3 的人口以水稻為主要食物來源(Brar and Khush 1997, Cho *et al*. 2004)。如何增加稻穀收穫量、改進稻米品質及增加其抗病蟲害和抗(耐)逆境等能力，是全球關心之課題。然而水稻品種很多，有的特別抗(耐)病蟲害及逆境，有些則否。對水稻之 EST 進行功能註解，有助於瞭解基因對病蟲害之防禦或逆境之調節的機制研究。褐飛蝨(brown planthopper, BPH) 屬同翅目飛蝨科 (*Homoptera* : *Delphacidae*)昆蟲，是亞洲稻作區分布最廣、危害最嚴重的水稻蟲害。自 Pathak *et al*. (1969)首次報導水稻品種對褐飛蝨具有抗性以來，利用水稻品種自身的抗性被認為是防治褐飛蝨危害的最佳途徑。抗性品種的選育與利用，一直是控制褐飛蝨種群最經濟、有效的措施(Chow 1992, Jeon *et al*. 1999, Wang *et al*. 1999, Zhao *et al*. 2005)。目前經由水稻 EST 功能註解所得之水稻對褐飛蝨抗性的基因功能表現之相關研究甚少，He *et al.* (2005) 利用抑制消減雜交法(suppression subtractive hybridization, SSH)以褐飛蝨取食 32 小時和未受褐飛蝨取食的水稻幼苗作為比對材料，建立了 SSH cDNA 基因庫，經反向總 RNA 墨點分析(reverse total RNA dot blot analysis)和北方墨點分析(Northern blot analysis)驗證，發現從 SSH cDNA 基因庫中所篩選的 25 個受褐飛蝨取食誘導之基因中有 17 個選殖與編碼已知蛋白質的基因有顯著的同源性，分別參與蛋白質的折疊、蛋白質與蛋白質間相互作用及訊息傳遞、物質傳輸和細胞增長等功能，其同時指出褐飛蝨取食啟動了水稻植株的衰老過程，且調節植物和昆蟲相互作用的訊息傳遞通路。水稻品系 B5 是一個對褐飛蝨抗性很強的品種，經多年多次抗褐飛蝨之鑑定中，均表現為基本不受褐飛蝨的危害(Yang *et al*. 1999)。由分子標記遺傳研究，已經證實 B5 於第 3 和第 4 染色體分別存在一個主效抗褐飛蝨基因(Hang *et al*. 2001, Wang *et al*. 2001)，不過它的抗蟲機制

目前還不清楚(Weng *et al*. 2003)。因此，為瞭解造成水稻基因對褐飛蝨抗性的功能機制，利用生物資訊學之方法，針對受褐飛蝨誘導之不同水稻抗、感性品種的 EST 進行功能註解之基因表現差異比較，有利於水稻分子遺傳研究。本研究藉由受褐飛蝨誘導之水稻 EST 序列資料為應用材料，建立一套快速且有效的 EST 功能註解分析流程。

材料與方法

一、試驗材料

本研究擷取 NCBI 之 dbEST 資料庫(Boguski *et al*. 1993, http://www.ncbi.nih.gov/dbEST/)中受褐飛蝨誘導之水稻EST序列共 188 條，這些 EST 分屬於感性水稻品種 Minghui 63 及抗性水稻品種 B5，將其下載後並分別儲存成兩個不同的 FASTA 格式之檔案，其相關資訊整理如 Table 1。這些序列都是由水稻苗期的組織所定序而來，長度介於 118~827 bp 之間，資料來源主要是由中國大陸所提供。

二、EST 的聚集與接合

由於不同的基因所表現的 mRNA 量不同，選殖時表現量多的基因較容易被選殖出來且產生較多數 EST 序列，表現量少的基因則較不易獲得 EST 序列。然而由於 EST 序列是隨機取樣之各轉錄產物 mRNA，表現量多的 EST 定序可能具有較高的重複性而導致產生多餘(redundant)序列，故可能會有多個 EST 代表同一個轉錄產物。因此在進行 EST 分析之前，為維持序列的品質，通常都會運用聚集(clustering)與接合(assembly)的方法，減少 EST 的重疊性，並將其連接成更長、更高質量的序列。這種分析之前的常規化處理對 EST 功能的識別、剪接產物的區分和基因表達圖譜的分析都有很大的幫助，進而有利於功能註解之進行(Adams *et al*. 1995)。聚集之目的就是將來自於同一個基因之具有重疊部分的 EST 歸至單一的群(cluster)，而無法歸群者則單獨成群，稱為 singleton 或 singlet，如此可建立一個非多餘性的序列資料庫(Kohchi *et al*. 1995, Bonaldo *et al*. 1996)；然後透過接合，即將同一群內的多個序列，經由剪切(clip) 5'和 3'端之低品質區域並計算讀取間的重疊性，再依重疊計分由大到小依序將讀取序列連結成較長的接合序列(contig 或 consensus sequence)。EST 經過聚集與接合所得的單獨成群之序列和接合序列，都屬於彼此之間沒有重疊的單一序列(unique sequences)。

CAP 序列接合程式，是一種 DNA 序列接合程式，能將短的讀取序列接合成長序列(Bonfield *et al*. 1995)，而 CAP3 (Huang and Madan 1999)是它的第三代版本，特別發展用來聚集和接合 EST 的一個成功程式，可獲得

Table 1. Information summary of rice ESTs induced by brown planthopper from NCBI-dbEST database (update time: May 31, 2005).

Genbank accession ID	Entry created	EST no.	Length (bp)	Cultivar[x]	Tissue	Country
BQ548081~BQ548082, BQ529294~BQ529324	2002/06/19	64	118~702	Minghui 63	seedling	China
BU673916~BU673930	2002/10/26					
CF568763~CF568780	2003/09/23					
BQ548083~BQ548087, BQ579829	2002/06/19	124	164~827	B5	seedling	China
BU572309~BU572426	2002/09/26					
Total		188	118~827			

[x] Minghui 63: susceptible variety; B5: resistant variety.

好品質的共用序列（Bhalerao 2003, Felipe *et al*. 2003）。故本研究利用 CAP3 序列接合程式對受褐飛蝨誘導之水稻 EST 序列，分別以分開抗、感品種之不同 cDNA 基因庫來進行聚集及接合。CAP3 序列接合程式可從 CAP3 官方網站(http://seq.cs.iastate.edu/)下載，需安裝至 Linux 作業平台上。CAP3 所讀取的序列檔案為 FASTA 格式，分析完成產生.ace 檔案格式的接合結果，包含有 seq.cap.contigs、seq.cap.contigs.qual 、 seq.cap.info 及 seq.cap.singlets 四個輸出檔案，分別儲放接合後所產生的接合序列內容、接合序列得分值、接合序列資料訊息及單獨成群序列內容。這些分析結果之檔案都能被儲存至 window 作業平台之個人電腦上以文字(text)形式來開啟，可供後續分析。本研究則是以 seq.cap.contigs 與 seq.cap.singlets 輸出檔案內所含的接合序列和單獨成群序列所組成的單一序列，進行後續的功能註解分析。

三、水稻 EST 功能註解分析流程之建立

本研究發現由美國基因組研究機構(The Institute for Genomic Research, TIGR, http://www.tigr.org)所建立的 TIGR Gene Index（TGI)資料庫（原由 TIGR 自己維護，從 2006 年底已移至 http://compbio.dfci.harvard.edu/tgi 網址，由美國的 Dana Farber 癌症研究所維護），擁有大量不同物種的 EST 資料，其中並包含有專門儲存已註解之水稻 EST 資料的 Rice（*Oryza sativa*）資料庫。TGI 蒐集的基因資料不僅是簡單地對 GenBank 的基因和 EST 數據按物種分類，也提取 GenBank 中不同物種的 EST 和註解的基因序列，並通過聚集、接合等分析過程（Sutton *et al*. 1995, Quackenbush *et al*. 2000, Lee *et al*. 2005）產生單一的、具有高可信度的假設性一致序列（tentative consensus sequence, TC），此 TC 可用來提供推測基因之功能註解，且除了可得知該 TC 序列本身的基本描述外，該網站亦提供可對應連結的其他相關之基因組訊息。於是，本研究將受褐飛蝨誘導之水稻 EST 所聚集接合的單一序列，直接利用 BLAST 去比對 TGI 資料庫，以進行相似性序列比對(Fig.1-1)；在所獲得之結果當中，若能比對到相似的 TC (Fig.1-2)，則發現每一條 TC 序列都有其相對應的 GO(Gene Ontology)分類代碼及敘述（Fig.1-3），然後由 GO 分類代碼（Fig.1-4, Fig.1-5）又能對應到 MIPS_FunCat 的分類代碼目錄（Fig. 1-6）。MIPS_FunCat 的分類目錄由 22 個主要功能種類所組成，可用來對不同生物之基因組進行註解之分類。而由基因本體聯盟(Gene Ontology Consortium)所維護的 GO (http://www.geneontology.org/)，主要提供結構及動態模式的控制字彙和分類系統，真核生物的基因和蛋白質也可透過 GO 的整合性分類系統得到分子功能(molecular function)、生物程序(biological process)、細胞成分(cellular component)三個層次的註解，可對基因組資訊進行簡單分類及描述（GO-EBI *et al*. 2006）。最後，我們便可以根據 GO 之三個層次分類及 MIPS_FunCat 之 22 個主要功能種類目錄，將比對到的結果，進一步進行基因各類功能表現量的統計分析。

因此，本研究將受褐飛蝨誘導之水稻 EST 單一序列的 FASTA 格式檔案，先以 BLAST 去比對 TGI 中 Rice（*Oryza sative*）之 TC 資料庫，該資料庫的搜尋比對之 E 值內設為 10，所謂 E 值定義為查詢序列與資料庫比對到的序列為隨機發生的期望值(機率)，E 愈小，表示兩序列真正有關係的機率愈高；由於本研究選定 Rice 之 TC 資料庫，所比對到的都會是水稻的 EST 單一序列，不可能包含著其他物種序列，因此一旦有比對到相似序列，只要由比對結果找出具最低 E 值的 TC，即代表查詢序列最有可能的功能註解。接著藉由 TC 搜尋結果連結到 GO 分類代碼及敘述，以進行功能註解之粗分類。然後經由 GO 分類代碼對應到 MIPS 分類代碼目錄，以

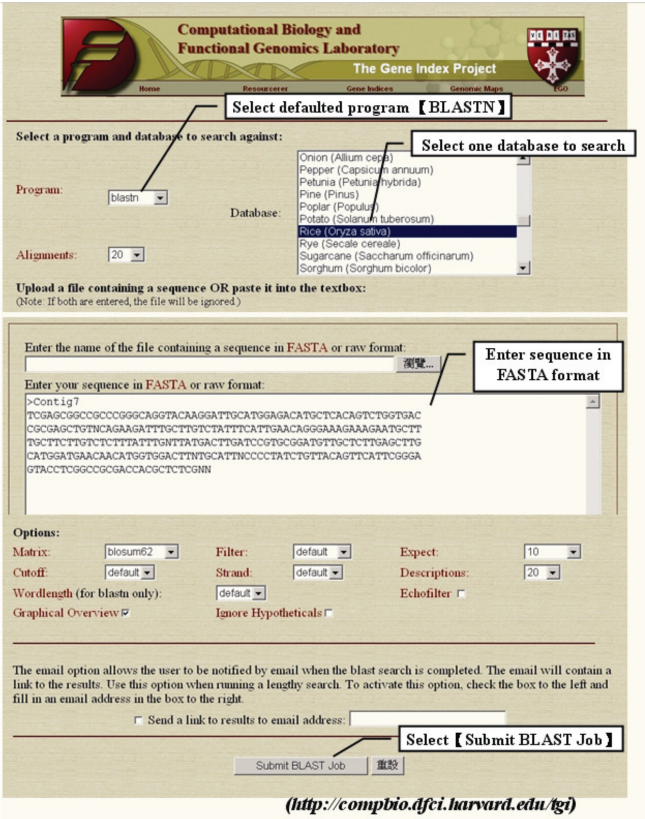

Fig.1-1. The step by step guide to EST functional annotation by searching TC database of TIGR Gene Index (TGI), and then linking the functional catalogs of GO and MIPS_FunCat: 【Step 1】 Sequence homology search by selecting program and database, and inputting the query sequences in TGI.

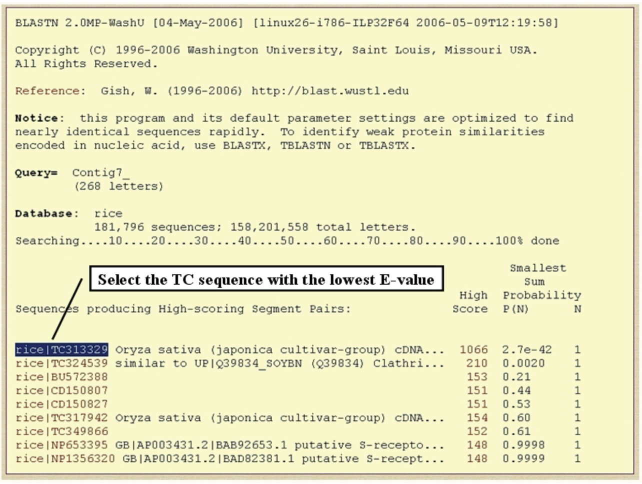

Fig. 1-2. The step by step guide to EST functional annotation by searching TC database of TIGR Gene Index (TGI), and then linking the functional catalogs of GO and MIPS_FunCat: 【Step 2】Select the TC sequence with the lowest E-value from the output result.

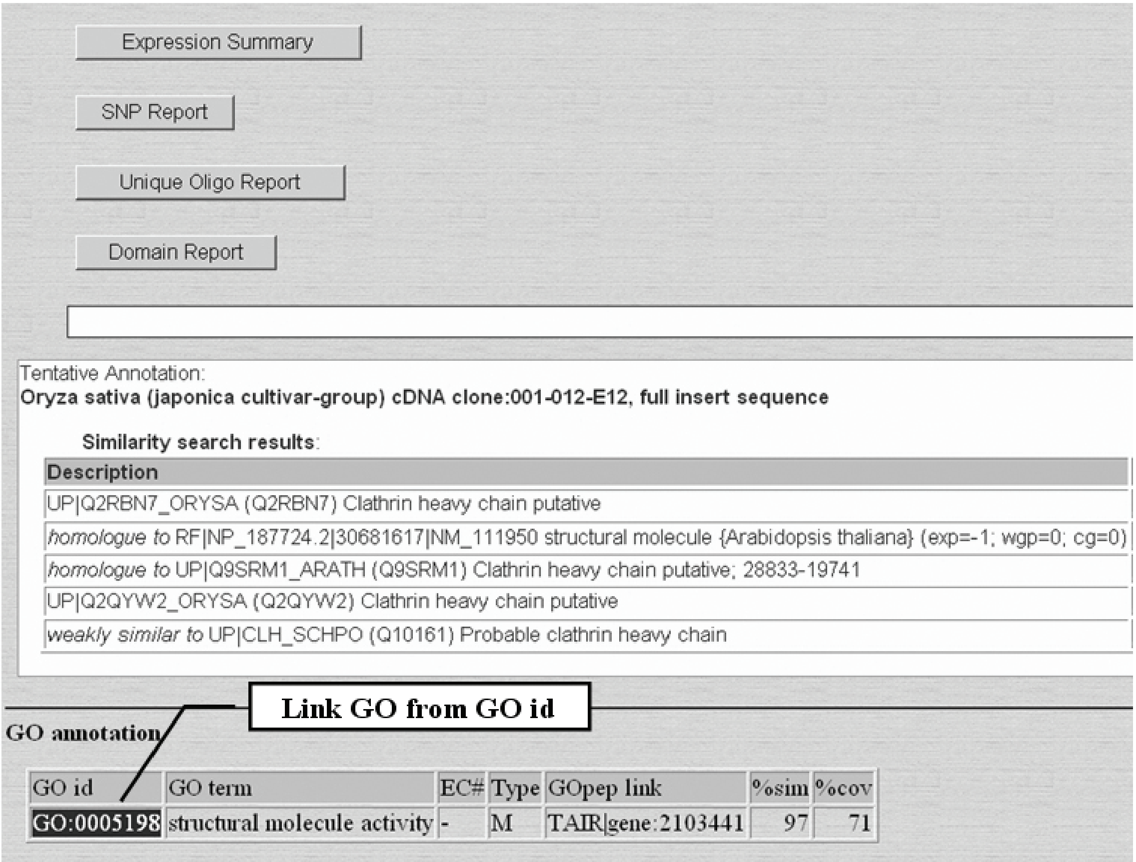

Fig. 1-3. The step by step guide to EST functional annotation by searching TC database of TIGR Gene Index (TGI), and then linking the functional catalogs of GO and MIPS_FunCat: 【Step 3】 Link GO from GO id of tentative annotation of the selected TC sequence.

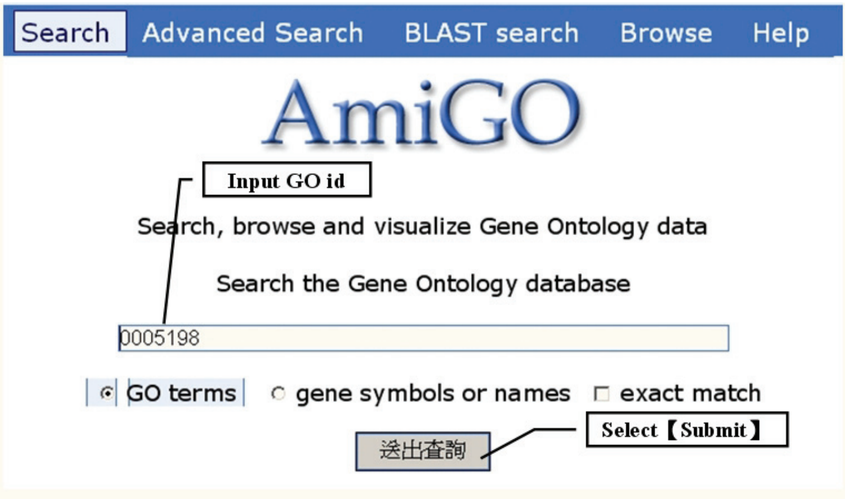

Fig. 1-4. The step by step guide to EST functional annotation by searching TC database of TIGR Gene Index (TGI), and then linking the functional catalogs of GO and MIPS_FunCat: 【Step 4】 Input the GO id of the selected TC sequence in GO browser (AmiGO).

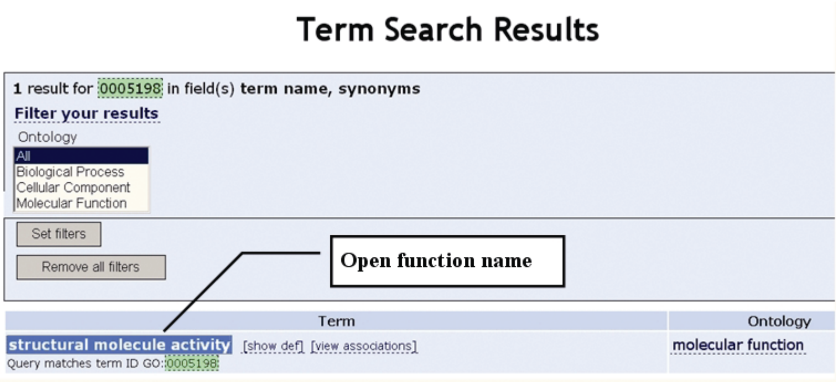

Fig. 1-5. The step by step guide to EST functional annotation by searching TC database of TIGR Gene Index (TGI), and then linking the functional catalogs of GO and MIPS_funcat:【Step 5】Open function name from term search results in AmiGO.

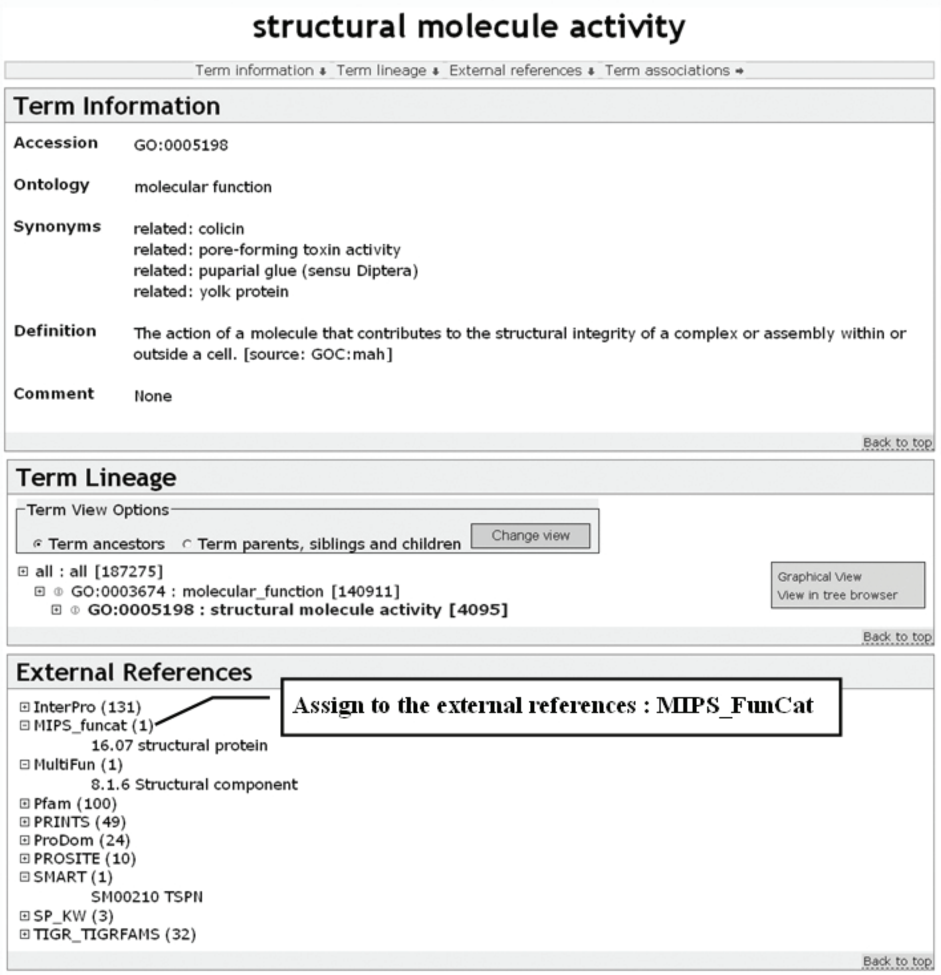

Fig. 1-6.The step by step guide to EST functional annotation by searching TC database of TIGR Gene Index (TGI), and then linking the functional catalogs of GO and MIPS_funcat:【Step 6】Link MIPS_FunCat from external references of GO id.

進行功能註解之細分類。最後，進行各類功能的個數統計及百分比計算。此功能註解分析流程，完整地包含了功能的註解、分類和表現量統計分析，如 Fig.2 所示。

結果

本研究擷取自 NCBI 之 dbEST 資料庫中受褐飛蝨誘導的水稻 EST 序列共 188 條，分屬於兩個品種，其中 Minghui 63 感性品種共 64 條，利用 CAP 3 序列接合程式進行聚集及接合後，共獲得 52 條較高品質和去除多餘性的單一序列，佔原有 EST 序列總數之 81%；B5 抗性品種中 EST 序列共 124 條，進行聚集及接合後，獲得共 103 條單一序列，佔原有 EST 序列資料之 83% (Table 2)。由此可知，經 CAP3 序列接合程式處理後，減少了 EST 所具有之序列的多餘性。

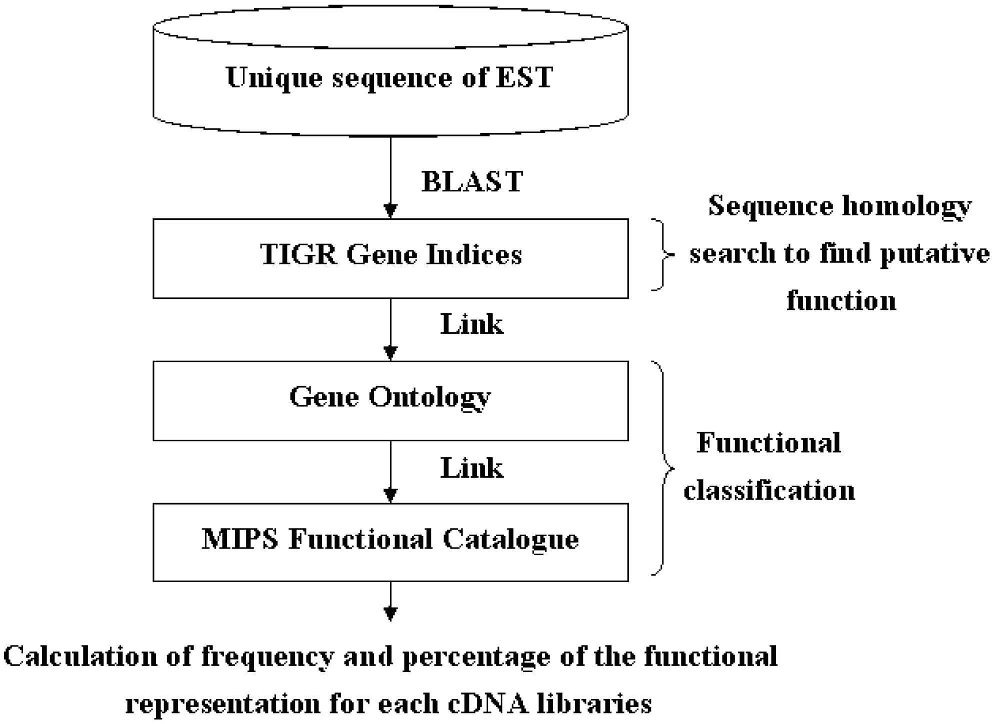

Fig. 2. Flowchart showing the analysis procedures of EST functional annotation established in this study.

Table 2. Clustering and assembly results with CAP3 software in rice ESTs induced by brown planthopper.

cDNA library	EST		Contig		Singleton			Unique sequence[x]	
	No.	Length (bp)	No.	Length (bp)	No.	%	Length (bp)	No.	%
Minghui 63	64	118~702	8	187~868	44	69	118~612	52	81
B5	124	164~827	10	237~877	93	75	177~827	103	83

[x] Unique sequences = contigs and singletons.

將受褐飛蝨誘導之水稻 EST 經聚集及接合後所產生的單一序列，依據本研究所建立之 EST 功能註解分析流程，最後比對搜尋所獲得之功能註解及分類結果，整理如 Table 3（此處限於篇幅，僅列 Minghui 63 的其中 15 筆為例說明）。

感性品種 Minghui 63 和抗性品種 B5 之 EST 單一序列，依據 GO 分類系統，所比對到的各功能類別的表現量及其所佔比例，如 Table 4 所示，可看出兩個抗、感品種之 EST 的功能在 GO 粗分類層次上之表現差異不大，均以生物程序(biological process)功能為主，其次是細胞組成(cellular component)功能，而分子功能(molecular function)相關基因則最少。

進而依據 MIPS 分類系統，來看更細分的各功能類別所佔表現量及其比例，如 Table 5 所示，在 MIPS 分類目錄的 22 個主要功能類別中，受褐飛蝨誘導之水稻 EST 單一序列僅比對到其中 17 項，其中兩個抗、感品種之 EST 功能表現都共同具有的功能有代謝(metabolism)、能量(energy)、細胞週期及 DNA 修飾(cell cycle and DNA processing)、蛋白質合成(protein synthesis)、蛋白質循環轉變(即蛋白質的折疊、修飾和傳送終點之生命週期，protein fate / folding, modification, destination)、細胞傳輸及其機制(cellular transport and transport mechanisms)、細胞間聯絡及訊息傳導之機制(cellular communication / signal transduction mechanism)、離子恆定性(ionic homeostasis)、系統性發育(development (systemic))、細胞組織之控制(control of cellular organization)、次細胞定位(subcellular localization)、蛋白質結合(protein binding)、傳輸促進作用(transport facilitation)。同時，Minghui 63 和 B5 兩品種之 EST 仍無法功能歸類之蛋白質(unclassified proteins)各佔 34.2% 和 27.5%。比較兩品種之基因功能表現差異，抗性品種(B5)啟動了感性品種(Minghui 63)所沒有的轉錄(transcription)、細胞逆境防禦反應(援救、防禦和毒性, cell rescue, defense and virulence)、細胞生長和形態發生(cell growth / morphogenesis)的基因功能表現。另外，感性品種在生物過程中的能量(energy)相關基因之表現量(13.1%)較抗性品種所得者(7.0%)高；而抗性品種在代謝(metabolism)相關基因之表現量(12.8%)則較感性品種所得者(3.6%)為高。

討論

本研究經由受褐飛蝨誘導之水稻 EST 資料，針對 EST 功能的註解、分類及表現量統計分析，成功建立了一套完整且快速的分析流程：(1)首先去比對 TGI 資料庫，以進行功能註解；(2)藉由搜尋結果連結至相對應之 GO 分類結果，可進行功能註解之粗分類；(3)再經由 GO 分類代碼連結至相對應之 MIPS 分類代碼目錄，以進行功能註解之細分類；(4)最後計算各類功能類別的個數統計及百分比計算。該流程不單是受褐飛蝨誘導之水稻 EST 資料，也適用於其他任何水稻 EST 資料的功能註解分析。而且只要在比對 TGI 資料庫時改選其他物種的資料庫，該分析流程一樣也能擴大應用對象至其他物種的 EST。

一般藉由 NCBI 資料庫進行相似性序列之搜尋比對，需耗費時間用來過濾與確認搜尋結果所獲得之相似性序列的物種來源。而在 MIPS 的 MAtDB 現有版本(2.0)資料庫內主要以阿拉伯芥基因組序列為主，水稻 EST 序列相關資料則較少，故目前利用此資料庫之相似性搜尋來對水稻 EST 序列進行功能註解，所能比對到的資訊會較少。由於 TGI 內有提供已註解之各物種的 EST 資料庫，因此本研究藉由 TGI 資料庫來對 EST 資料進行功能註解，可節省過濾篩選序列所需的大量人力及時間。另外，透過 TGI 還可得知其他相對應之基因組訊息，例如開放讀架(open reading frame, ORF)、單核酸多型性 (single

Table 3. The results of functional annotation and classification for rice ESTs induced by brown planthopper (only display the partial 15 sequences of Minghui 63).

Unique sequence		TC	E-value	Tentative annotation	GO id	GO category [x]	MIPS _FunCat [y]
ID	Length (bp)						
contig 1	514	TC268557	3.6×10^{-91}	similar to Drosophila mRNA to mitochondrial 16S ribosomal RNA gene, complete	-	-	-
contig 2	318	TC263601	2.7×10^{-57}	(Q8W2X7) Putative ubiquitin-conjugating enzyme, complete	GO:0006464 GO:0006511 GO:0005515 GO:0005234	P P F C	6 6 63 40
contig 3	670	TC251056	1.9×10^{-138}	similar to (Q75I68) Hypothetical protein OSJNBa0083F15.14, complete	-	-	-
contig 4	344	TC260993	6.6×10^{-57}	(Q6ZL95) Putative photosystem I antenna protein, complete	GO:0015979	P	2
contig 5	458	TC250268	8.1×10^{-72}	similar to (Q8RZH3) Putative CTP synthase, partial (92%)	GO:0006118	P	2
contig 6	718	TC261354	2.8×10^{-23}	-	-	-	-
contig 7	883	TC251363 TC262426 TC251362	2.8×10^{-126} 3.9×10^{-111} 1.6×10^{-108}	(Q6K9S2) Putative brown planthopper-induced resistance protein 1, complete; (Q6V9H6) Brown planthopper-induced resistance protein 1, partial (73%); similar to (Q93VK5) At1g31800/68069_m00159, partial (77%)	-	-	-
contig 8	868	TC262303	8.7×10^{-147}	(Q6K9S2) Putative brown planthopper-induced resistance protein 1, partial (85%)	-	-	-
BQ548083	461	TC263805	5.3×10^{-81}	(Q6Z345) Dof zinc finger protein, complete	-	-	-
BQ548084	447	TC248131	8.2×10^{-90}	(Q6Z246) Putative signal recognition particle receptor, complete	GO:0006605	P	6
BQ579829	307	TC262512	1.8×10^{-55}	(Q7X7H3) OSJNBa0084K20.14 protein (OSJNBa0076N16.12 protein), complete	GO:0009507	C	40
BU572309	401	TC249638	6.8×10^{-74}	similar to (P10862) Postreplication repair protein RAD18, partial (5%)	-	-	-
BU572310	528	TC252310	2.4×10^{-95}	(Q6ZL26) Mitogen activated protein kinase kinase, complete	GO:0006468	P	6
BU572312	656	TC261354	4.8×10^{-08}	-	-	-	-
BU572315	561	TC261348	6.4×10^{-18}	immunoglobulin heavy chain variable region {Callithrix jacchus;}, partial (5%)	-	-	-

[x] F, molecular function; P, biological process; C, cellular component .

[y] 1, metabolism; 2, energy; 3, cell cycle and DNA processing; 4, transcription; 5, protein synthesis; 6, protein fate (folding, modification, destination); 8, cellular transport and transport mechanisms; 10, cellular communication / signal transduction mechanism; 11, cell rescue, defense and virulence; 13, ionic homeostasis; 14, cell growth / morphogenesis; 20, Plant hormonal regulation; 25, development (systemic); 29, Transposable elements, viral and plasmid proteins; 30, control of cellular organization; 40, subcellular localization; 62, Regulator of G-protein signaling; 63, protein binding; 65, Storage protein 67, transport facilitation; 98, Classification not yet clear-cut; and 99, unclassified proteins.

Table 4. Statistics of the gene functional representation for rice ESTs induced by brown planthopper, based on GO functional category.

cDNA library	GO functional category					
	Molecular function		Biological process		Cellular component	
	No.	%	No.	%	No.	%
Minghui 63	26	31.5	35	38.0	31	33.7
B5	74	28.8	100	38.9	83	32.3

Table 5. Statistics of the gene functional representation for rice ESTs induced by brown planthopper, based on MIPS functional category.

MIPS functional category		Minghui 63		B5	
ID	Definition	No.	%	No.	%
1	Metabolism	7	3.6	18	12.8
2	Energy	10	13.1	12	7.0
3	Cell cycle and DNA processing	2	1.2	2	1.2
4	Transcription	0	0.0	1	0.4
5	Protein synthesis	1	2.2	13	4.2
6	Protein fate (folding, modification, destination)	2	3.9	11	12.4
8	Cellular transport and transport mechanisms	2	2.2	7	3.5
10	Cellular communication/ signal transduction mechanism	2	1.2	3	0.5
11	Cell rescue, defense and virulence	0	0.0	1	0.4
13	Ionic homeostasis	1	1.1	2	0.7
14	Cell growth/ morphogenesis	0	0.0	4	0.4
25	Development (Systemic)	3	1.3	3	0.4
30	Control of cellular organization	3	2.8	5	2.5
40	Subcellular localization	15	13.6	39	18.6
63	Protein binding	4	8.4	7	3.5
67	Transport facilitation	4	11.1	4	3.8
99	Unclassified proteins	11	34.2	26	27.5

nucleotide polymorphism, SNP)、GO 註解等，此有利於 EST 功能註解分析之工作。MIPS_FunCat雖然可提供線上搜尋之功能分類，但是它是以逐筆輸入關鍵字的方式來進行搜尋而得到分類結果；此結果會因使用者輸入的關鍵字不同而有差異，往往造成同一筆資料，卻對應到多個不同的分類結果，讓使用者很難確切得知真正分類結果到底為何，相對地也影響了資料分析之準確度。另外，由於 MIPS_FunCat 所提供之搜尋方法是一次僅能對單一筆資料來進行搜尋分析，這對少量資料而言不成問題，不過一旦分析資料量達上萬筆甚至數十萬筆時，需要花費的工程相當龐大，是不合實際的。因此，本研究所建立之 EST 功能註解分析流程，先透過比對 TGI 資料庫，再藉由搜尋結果連結至相對應之 GO 分類結果，然後經由 GO 分類結果連結至相對應之 MIPS 分類目錄，以解決此問題，無須再擔心因輸入關鍵字不當或不夠而影響分析之準確度。此對於未來各研究單位要對各物種 EST 資料進行功能註解分析研究時，有很大之幫助。再者，由於本研究所建立的功能註解分析流程是從 TGI 對應到 GO 再對應到 MIPS，這三個程序之間存在有一定且直接的代碼對應關係，因此可進一步藉此將相關資料(TC 序列本身及其相關資

訊、GO 和 MIPS 分類代碼)下載至本端主機，以便能將整個分析流程建立起一個 EST 功能註解的自動化分析系統，可有利於處理巨量 EST 的功能註解(Lu *et al*. 2006)。至於直接去比對 NCBI 進行 EST 功能註解然後再去 MIPS_FunCat 做功能分類的傳統方式，在 NCBI 和 MIPS_FunCat 之間由於欠缺對應關聯性，難以將其整合成能一氣呵成的自動化程序。若採以比對 MIPS_MAtDB 後再以 MIPS_FunCat 進行 EST 功能註解和分類的傳統方式，受限於目前該網站只提供阿拉伯芥基因組序列之下載而已，致無法在本端主機建構一個適合於水稻 EST 功能註解分析的自動化系統。

褐飛蝨於臺灣第二期稻作乳熟期至糊熟期間危害最嚴重，卵產於葉鞘脊部組織內，成蟲與若蟲均好群集於稻株基部吸取稻液為害，稻株受害輕者莖葉黃化，嚴重時引起全株枯死而造成「蝨燒」現象。這是因為褐飛蝨口針插入水稻韌皮部組織，吸取韌皮部汁液作為食物來源，其取食後會嚴重干擾水稻的正常生理活動、降低光合作用速率和干擾光合產物的分配(Watanabe and Kitagawa 2000, Yuan *et al*. 2005)。由於昆蟲的取食作用會刺激活化植株一系列防禦基因的表達，產生誘導型抗性以抵禦植食性昆蟲的進一步危害(Kessler and Baldwin 2002, Yuan *et al*. 2005)，因此了解水稻與褐飛蝨互相作用的機制原理，將有助於培育抗蟲水稻品種，以及研發對環境無污染的殺蟲劑(Frutos *et al*. 1999, Weng *et al*. 2003)。國內外普遍認為抗、感植株二次代謝物質的差異，是水稻品種抗感與否的主要原因(Chow 1992, Kong *et al*. 2002, Zhao *et al*. 2005)。Scheideler *et al*. (2002) 以阿拉伯芥(*Arabidopsis*)為材料，指出細胞的逆境防禦反應（援救、防禦及毒性)、訊息傳遞和經由增強代謝路徑可抵抗不同壓力之傷害。而由本研究結果顯示，受褐飛蝨誘導之水稻抗性品種的 EST 功能以代謝、轉錄、和細胞逆境防禦反應為主，因此可知基因大量表現出代謝、轉錄和細胞逆境防禦反應之功能，是使水稻產生抗性來抵抗病原菌之侵襲或環境壓力等傷害的主要因素，此顯示水稻受褐飛蝨誘導後會刺激活化水稻的防禦反應，同時也表示出植物抗蟲害反應涉及到複雜的訊息傳遞，如此高含量的訊息傳遞因子的存在，可以即時對內外環境的變化作出回應，以調整有機體狀態來適應環境的變化。抗性品種也啟動了感性品種所沒有的轉錄功能，表示這部份基因在植物防禦反應中也是不可缺少的。而與細胞生長和形態發生相關的基因在抗性品種的表現含量較高，可能的原因是當植物進行防禦反應過程時，細胞生長、發育過程受到刺激活化，促使有機體能更強壯以增加防禦的能力；反之，感性品種在受到褐飛蝨侵害時細胞生長、發育過程受到暫時性抑制。另外，本研究結果發現受褐飛蝨誘導之水稻感性品種 EST 之能量功能的表現量高於抗性品種，這是因為水稻在防禦過程中需要極大的能量消耗，而感性品種的代謝分解能力較缺乏，因此當它表現代謝作用的同時，所需要消耗的能量也會較高，具有能量功能的基因也因而大量表現。Hei *et al*. (2005)指出受褐飛蝨取食誘導之基因參與核酸合成、細胞防禦、訊息傳遞等功能，且防禦反應的基因在褐飛蝨取食後表達會增強，對此本研究結果也印證了受褐飛蝨誘導之水稻抗性品種能表現出感性品種所沒有的防禦功能。由此可知，藉由受褐飛蝨誘導之水稻 EST 功能註解分析對有功能基因的研究發現，不但可以解釋水稻在受褐飛蝨侵害過程中各種生理生化反應的變化狀況，也有利於瞭解水稻對蟲害之基因防禦機制的研究。從本研究所發現的抗、感性品種之間存在的共有基因功能表現，表示水稻在受褐飛蝨誘導後在某些生理生化反應過程中仍存在一定的共同性，即一些相關蛋白都在水稻防禦褐飛蝨侵害過程中產生一定的作用，此亦證實水稻的抗蟲過程是涉及眾多基因共同表現的一個複雜的網絡體系。而

抗、感性水稻品種之間存在的特有基因功能表現，可闡明造成水稻對褐飛蝨抗性或感性的主要原因，乃涉及到複雜的轉錄、細胞逆境防禦反應、細胞生長和形態之相關基因。在長期的自然選擇和物種競爭中，抗性品種形成了一套適應的防禦體系，由於具有這些高度表現的基因，因此增強了水稻對褐飛蝨的防禦能力。

引用文獻

Adams MD, M Dubnick, JM Kelley, JC Venter, *et al*. (1992) Sequence identification of 2,375 human brain genes. **Nature** 355:632-634.

Adams MD, JM Kelley, JD Gocayne, M Dubnick, MH Polymeropoulos, H Xiao, CR Merril, A Wu, B Olde, RF Moreno, *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. **Science** 252:1651-1656.

Adams M, AR Kerlavage, RD Fleischmann, *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. **Nature** 377:173-174.

Adams M, JC Venter, *et al*. (1993) Chromosomal distribution of 320 genes from a brain cDNA library. **Nat. Genet.** 4(4):381-386.

Andreas R, A Zollner, D Maier, J Hani, HW Mewws, *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. **Nucleic Acids Res.** 32(18):5539-5545.

Bhalerao R (2003) Gene Finding in *Populus* – the Bioinformatics of a EST Program. Department of Biotechnology, Royal Institute of Technology, Stcokholm, Sweden. Umeå. 44pp.

Boguski MS, TM Lowe, CM Tolstoshev (1993) dbEST--database for "expressed sequence tags". **Nat. Genet**. 4 (4): 332-333.

Bonaldo MF, G Lennon, *et al*. (1996) Normalization and subtraction: two approaches to facilitate gene discovery. **Genome Res**. 6(9):791-806.

Bonfield JK, K Smith, R Staden (1995) A new DNA sequence assembly program. **Nucleic Acids Res**. 24:4992–4999.

Brar DS, GS Khush (1997) Alien introgression in rice. **Plant Mol. Biol.** 35:35-47.

Brenner WD, MS Katari, DW Stevenson, SA Rudd, AW Douglas, WN Moss, RW Twigg, SJ Runko, GM Stellari, WR McCombie, GM Coruzzi (2005) EST analysis in Ginkgo biloba: an assessment of conserved developmental regulators and gymnosperm specific genes. **BMC Genomics** 6:143.

Cho, SK, SH Ok, JU Jeung, KS Shim, *et al*. (2004) Comparative analysis of 5211 leaf ESTs of wild rice (*Oryza minuta*). **Plant Cell Reports** 22:839-847.

Chow M (1992) Principles and Applications of Insect Resistance in Crops. (in Chinese) Peking Agric. Unvi. Press. 359pp.

Felipe MS, RV Andrade, SS Petrofeza, AQ Maranhao, FA Torres, *et al.* (2003) Transcriptome characterization of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis* by EST analysis. **Yeast** 20(3):263-271.

Frishman D, K Albermann, J Hani, *et al.* (2001) Functional and structure genomics using PEDANT. **Bioinformatics** 17:44-57.

Frutos R, C Rang, M Royer (1999) Managing insect resistance to plants producing *Bacillus thuringiensis* toxins. **Critical Rev. Biotechnol.** 19(3):227-276.

GO-EBI, EMBL-EBI, *et al*. (2006) The Gene Ontology (GO) project in 2006. **Nucleic Acids Res.** 34:322-326.

Hang Z, G He, L Shu, *et al.* (2001) Identification and mapping of two brown planthopper genes in rice. **Theor. Appl. Genet.** 102:929-934.

He Q, HY Yuan (2005) The construction and screening of suppression subtractive cDNA library of rice seedling exposed to BPH. (in Chinese) **Guihaia** 25(3):237-240.

Huang X, A Madan (1999) CAP3: A DNA sequence assembly program. **Genome Res.** 9:868-877.

Jeon YH, SN Ahn, H Chune, *et al.* (1999)

Identification of a RAPD marker linked to a brown planthopper resistance gene in rice. **Euphytica** 107:23-28.

Jongeneel CV (2000) Searching the expressed sequence tag (EST) databases: panning for genes. **Brief. Bioinform.** 1:76-92.

Kantety RV, ML Rota, DE Matthews, *et al.* (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. **Plant Mol. Biol.** 48:501-510.

Kawai J, A Shinagawa, K Shibata, M Yoshino, M Itoh, Y Ishii, T Arakawa, *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. **Nature** 409: 685-690.

Kelly JM, M Adams, *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. **Science** 252:1651-1656.

Kessler A, IT Baldwin (2002) Plant responses to insect herbivory: the emerging molecular analysis. **Ann. Rev. Plant Biol.** 53:299-328.

Kohchi T, K Fujishige, *et al.* (1995) Construction of an equalized cDNA library from *Arabidopsis thaliana*. **Plant J.** 8(5):771-776.

Kong CH, XH Xu, F Hu, XH Chen, X Chen, B Ling, Z Tan (2002) Using specific secondary metabolites as markers to evaluate allelopathic potentials of rice varieties and individual plants. **Chinese Sci. Bull.** 47(3):203-206.

Lai J, N Dey, CS Kim, AK Bharti, S Rudd, KFX Mayer, BA Larkins, P Becraft, J Messing (2004) Characterization of the maize endosperm transcriptome and its comparison to the rice genome. **Genome Res.** 14:1932-1937.

Lee Y, J Tsai, S Sunkara, S Karamycheva, G Pertea, R Sultana, V Antonescu, A Chan, F Cheung, J Quackenbush (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. **Nucleic Acids Res.** 33:71-74.

Lu HY, CT Chen, YC Chiu, CT Lu, ML Wei (2006) An automated system for EST functional analysis and its application in rice genes. p.77-78. *In:* Bioinformatics in Taiwan Symposium and Workshop, 2006. Taichung, Taiwan.

Mewes HW, D Frishman, KFX Mayer, O Noubibou, *et al.* (2006) MIPS : analysis and annotation of proteins from whole genomes in 2005. **Nucleic Acids Res.** 34:169-172.

Pathak MD, CH Cheng, ME Fortuno (1969) Resistance to *Nephotettis impicticeps* and *Nilavarpata lugens* in varieties of rice. **Nature** 223:502-504.

Quackenbush J, F Liang, I Holt, G Pertea, J Upton (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. **Nucleic Acids Res.** 28(1):141-145.

Rounsley SD, A Glodek, G Sutton, MD Adams, *et al.* (1996) The construction of *Arabidopsis* expressed sequence tag assemblies. A new resource to facilitate gene identification. **Plant Physiol.** 112(3):1177-1183.

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequence? **Trends Plant Sci.** 8:321-329.

Scheideler M, NL Schlaich, K Fellenberg, T Hauser, NC Vingron, *et al.* (2002) Monitoring the switch from housekeeping to pathogen defense metabolism in *Arabidopsis thaliana* using cDNA arrays. **J. Biol. Chem.** 277:10555-10561.

Shim KS, SK Cho, JU Jeung, KW Jung, *et al.* (2004) Identification of fungal (*Magnaporthe grisa*) stree-induced genes in wild rice (*Oryza minuta*). **Plant Cell Rep.** 22:599-607.

Sreenivasuls N, PB Kavi Kishor, RK Varshney, L Altschmied (2002) Mining functional information from cereal genomes – the utility of expressed sequence tags. **Curr. Sci.** 83:965-973.

Sutton G, O White, M Adams, *et al.* (1995) TIGR Assembler: a new tool for assembling large shotgun sequencing projects. **Genome Sci. Technol.** 1:9-19.

Varshney RK, T Thiel, N Stein, P Langridge, *et al.* (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. **Cellular Mol. Biol. Letters** 7:537-546.

Wang J, X Yu, Z Lu, *et al.* (1999) Breeding of resistance to brown planthopper (*Nilaparvata*

lugens) in indica hybrid rice. (in Chinese) **Chinese J. Rice Sci.** 13(4):242-244.

Wang B, Z Huang, L Shum, X Ren, X Li, G He (2001) Mapping of two new brown planthopper resistance genes from wild rice. **Chinese Sci. Bull.** 46(13):1092-1096.

Watanabe T, H Kitagawa (2000) Photosynthesis and translocation of assimilates in rice plants for following phloem feeding by the planthopper *Nilaparvata lugens*. **J. Econom. Entomol.** 93:1192-1198.

Wei ML, CT Lu, HY, Lu (2005a) Research and development of expressed sequence tag (EST) in cereal crops. (in Chinese) **Sci. Agric.** 53(1,2):15-21.

Wei ML, HY Lu, CT Lu (2005b) Bioinformatics collection and statistics of rice disease/pest resistance genomics research. (in Chinese) **Crop Environ, Bioinform.** 2(4):295-306.

Weng QM, Z Huang, XL Wang, LL Zhu, GC He (2003) *In situ* localization of proteinase inhibitor mRNA in rice plant challenged with brown planthopper. **Chinese Sci. Bull.** 48:979-982.

Wolfsberg TG, D Landsman (2001) Expressed sequence tags (ESTs). p.283-301. *In:* Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. 2*nd* ed. DB Andreas, BFF Ouellette (eds.) John Wiley & Sons, Inc. NY.

Yang CJ, ZH Yang, Hu JF, He GC, LH Shu (1999) Study on the brown planthopper resistance in introgressive lines from wild rice. (in Chinese) **Acta Phytophylacica Sinica** 26:197-202.

Yuan HY, LL Zhu, GC He (2005) Construction and screening of a cDNA library from rice seedling exposed to brown planthopper. (in Chinese) **J. Xinyang Normal Univ. (Nat. Sci. Ed.)** 18(1):48-51.

Zhao Y, FK Huang, XL Tong (2005) Content variations of secondary compounds in rice plants and their influence on rice resistance to brown planthopper, *Nilaparvata lugens*. (in Chinese) **Chinese J. Rice Sci.** 19(5):479-482.

－編輯：楊純明